SLLIM: System Log Local Intelligent Model

Authors: Carlos Cruzportillo, Nassos Galiopoulos, Jason Gillette

Affiliation: University of Texas at San Antonio

Date: October 7th, 2024

Introduction

In today’s digital age, the widespread adoption of internet-connected devices—ranging from IoT devices to mobile phones—has created an unprecedented volume of data. Enterprises, relying on complex systems, face a growing challenge in monitoring, managing, and securing system-generated logs.

Problem Statement

The increasing volume of system logs generated by interconnected devices and enterprise systems creates a challenge for IT professionals in efficiently detecting threats and diagnosing issues, necessitating the development of lightweight, intelligent tools for real-time log analysis and query.

Research Questions

1. How well can lightweight LLMs detect system issues and security threats from system logs?
2. How effectively can lightweight LLMs perform question answering compared to larger, more resource-intensive models?

Specific Objectives

1. Fine-tune at least two lightweight LLMs for comparative analysis.
2. Evaluate question answering performance of lightweight LLMs versus resource-intensive models in the cybersecurity domain.

Literature Review

1. **[Paper 1]**: Summary and relevance
2. **[Paper 2]**: Summary and relevance
3. **[Paper 3]**: Summary and relevance
4. **[Paper 4]**: Summary and relevance
5. **[Paper 5]**: Summary and relevance

Paper 1

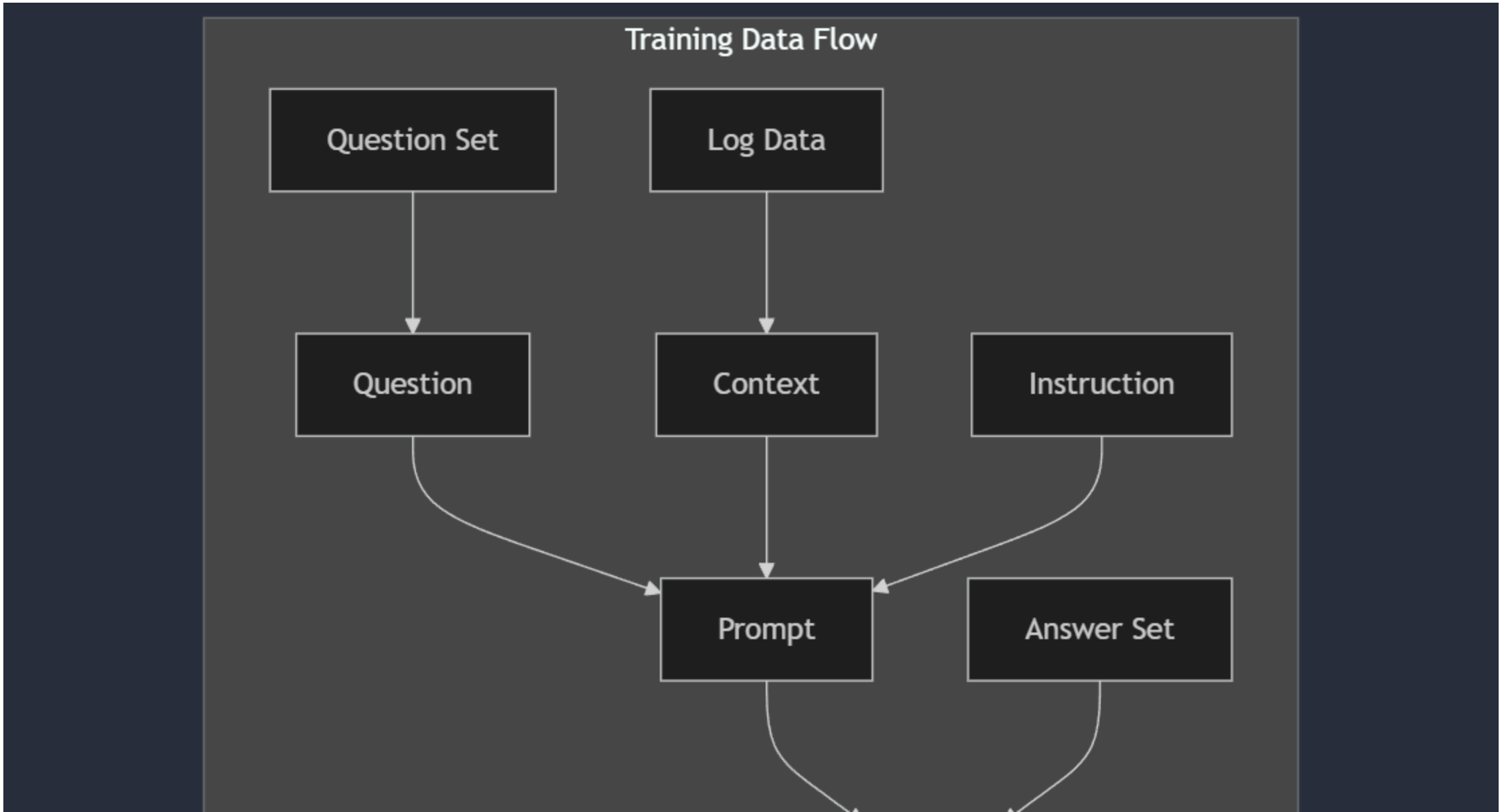
Paper 2

Paper 3

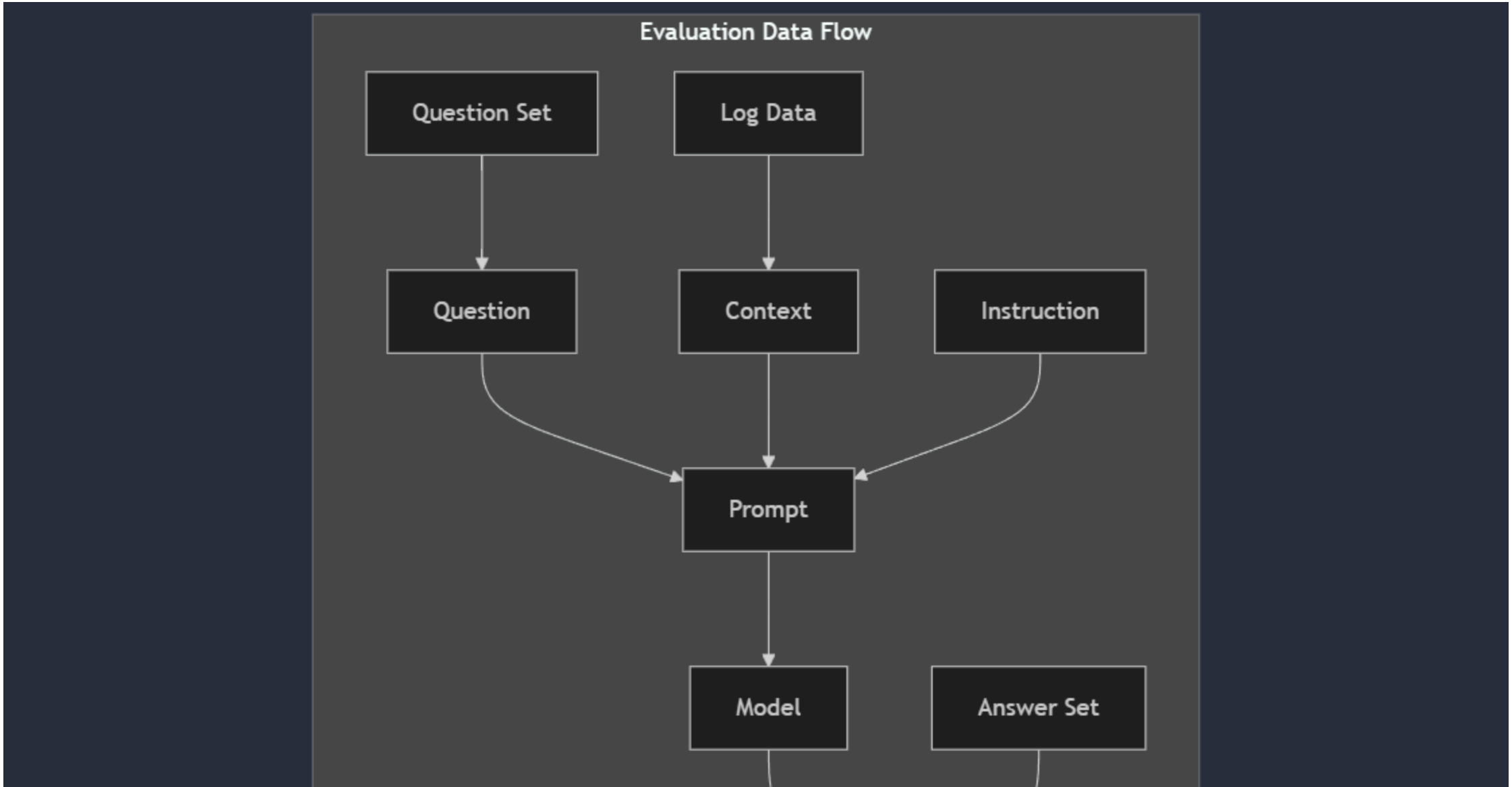
Methodology: Lightweight LLMs

Lightweight models perform tasks such as question answering on system logs while minimizing resource consumption. Techniques like **model compression**, **knowledge distillation**, **pruning**, and **efficient layer design** ensure that these models run efficiently in resource-constrained environments.

Question) pattern. This process allows the model to learn how to generate accurate responses to system log queries.



match the expected answers based on contextual embeddings, providing a robust metric for tasks where meaning is more important than exact word matching.



Datasets

1. **Dataset 1**: [Placeholder]
2. **Dataset 2**: [Placeholder]

Expected Outcomes

We hypothesize that lightweight LLMs will deliver faster inference times and consume fewer computational resources while maintaining competitive accuracy in domain-specific tasks like system log analysis, making them more efficient than larger models in this context.

Project Timeline

Milestone	Deadline
Proposal Submission	20241007
Dataset Collection	20241014
Model Training	TBD
Evaluation and Testing	TBD
Final Report Submission	20241202

Challenges and Risks

1. **Data Quality**: Ensuring question-answer pairs are aligned with log data context in each prompt.
2. **Data Preparation**: Chunking or retrieving relevant log subsets as context without losing essential information.
3. **Model Overfitting**: Avoiding overfitting to domain-specific logs, which may hinder generalization to unseen questions.

Mitigation Strategies

1. **Data Quality**: Implement strict human-in-the-loop validation to cross-check question-answer pairs against log data.
2. **Data Preparation**: Develop an automated log chunking and retrieval pipeline using retrieval tools to ensure full context is covered.
3. **Model Overfitting**: Use techniques like cross-validation, regularization, and dropout, and incorporate diverse training examples to maintain generalization.

References

1. [Placeholder for Reference 1]
2. [Placeholder for Reference 2]
3. [Placeholder for Reference 3]
4. [Placeholder for Reference 4]
5. [Placeholder for Reference 5]